PROTEINS: Structure, Function, and Genetics 19:199–221 (1994)

HOOK: A Program for Finding Novel Molecular Architectures That Satisfy the Chemical and Steric Requirements of a Macromolecule Binding Site

Michael B. Eisen,[1,3] Don C. Wiley,[2,3] Martin Karplus,[4] and Roderick E. Hubbard[5]
[1]Committee on Higher Degrees in Biophysics, [2]Howard Hughes Medical Institute, [3]Department of Biochemistry and Molecular Biology and [4]Department of Chemistry, Harvard University, Cambridge, Massachusetts 02138 and [5]Department of Chemistry, University of York, Heslington, York, England YO1 5DD

ABSTRACT **A program (HOOK) is described for generating potential ligands that satisfy the chemical and steric requirements of the binding region of a macromolecule. Functional group sites with defined positions and orientations are derived from known ligand structures or the multicopy simulation search (MCSS) method (Miranker, A., Karplus, M. Proteins 11:29–34, 1991). HOOK places molecular "skeletons" from a database into the protein binding region by making bonds between sites ("hooks") on the skeleton and functional groups. The nonpolar interactions with the binding region of candidate molecules are assessed by use of a simplified van der Waals potential. The method is illustrated by constructing ligands for the sialic acid binding site of the hemagglutinin from the influenza A virus and the active site of chloramphenicol acetyltransferase. Aspects of the HOOK program that lead to a highly efficient search of 10^5 or more skeletons for binding to 10^2 or more functional group minima are outlined.**

Key words: multicopy simulation search, rational drug design, database search, computer-aided design

INTRODUCTION

Drug design is concerned with the generation of ligands that bind strongly to key regions of biologically important molecules (e.g., enzyme active sites, receptor proteins) so as to inhibit or alter their activity. An immense effort has been and continues to be dedicated to developing methods by which drug design (or more properly, ligand design, since whether or not a ligand is a drug involves factors beyond present concerns) can be made a more rational process. However, the field is still in its infancy as is evident from recent reviews,[1] as well as from earlier evaluations.[2,3] Thus, new ideas and methods for approaching the ligand design problem are still needed.

The strategy for rational ligand design that is being developed has three parts. The first is an efficient method for the search of unknown binding regions (or more generally receptor surfaces) for locations where a range of functional groups can interact strongly with the protein. Second, given a set of such positions and orientations for functional groups, it is necessary to connect the functional groups to form molecules that are candidates for synthesis. Finally, a method is needed to determine which of the resulting molecules are likely to have the strongest binding constants. A stepwise procedure is more efficient than doing everything at once, i.e., the first two steps, which need to be as inclusive as possible, can make use of relatively simple representations of the interactions; evaluating the resulting candidates in the third step requires a more sophisticated and time consuming treatment of the interactions that can be applied only to a limited set of molecules.

To solve the first problem, the multicopy simultaneous search (MCSS) method was developed.[4] It is based on a combination of random placement and energy minimization/quenched dynamics techniques that make possible an efficient determination of energetically favorable positions and orientations of functional groups in the binding site of a protein whose three-dimensional structure is known. Functional groups are small chemical fragments commonly found as substituents of larger organic molecules (see Fig. 1a for examples). The method begins by filling a binding site with multiple copies of a functional group. The positions and orientations of these groups are minimized against a fixed model of the binding site until local energy minima are found. The energy function that is minimized includes only the interactions between individual copies of the functional group and the protein (functional groups see only the protein, and not each

Received June 18, 1993; revision accepted January 28, 1994.
Address reprint requests to Martin Karplus, Department of Chemistry, Harvard University, 12 Oxford Street, Cambridge, MA 02138.

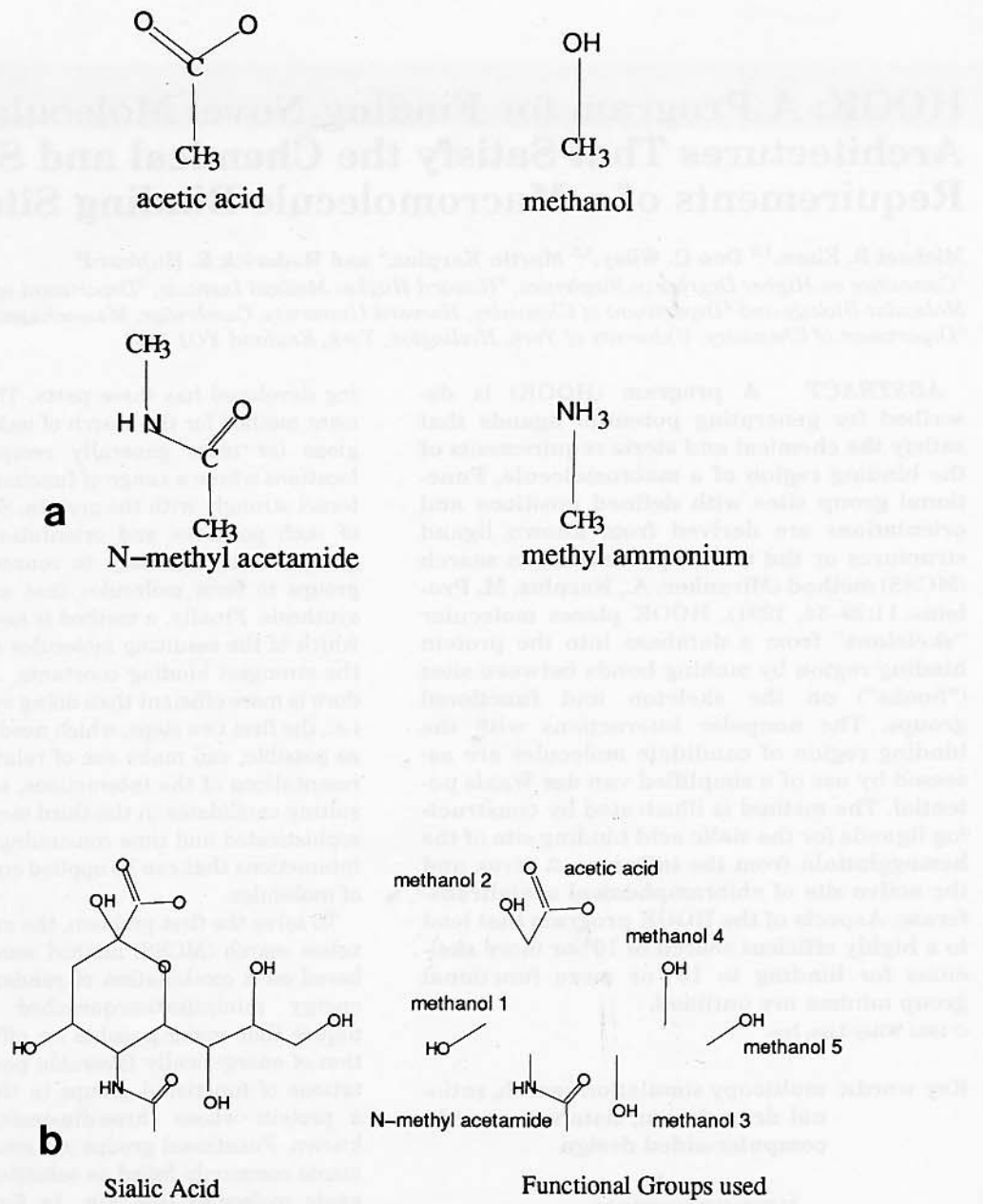


Fig. 1. (**a**) Functional groups used in the MCSS calculations discussed in this paper. The groups are all rigid, and have a free CH_3 group. (**b**) The sialic acid molecule highlighting the seven functional group sites. The methanol groups are labeled 1–5 for reference.

other), so that the resulting output is a collection of discrete sites (called functional group sites) where functional groups are likely to make favorable interactions with the protein. The collection of these functional group sites, which include both a position and orientation for each site, is known as a functionality map.

The MCSS method supplements the GRID program,[5,6] which is the best known and most widely used method for functional group searches. GRID determines favorable binding sites based on an interaction grid determined by an empirical energy function. It uses simple spherical ligands and is limited to obtaining positional information. Of special importance in the MCSS method is its use of chemically realistic functional groups that permit determination of their expected orientation, as well as their positions. In particular, it provides CH_3–R bond vectors that can be used to join them together to form candidate ligands. Also, the MCSS method

with a quenched dynamics protocol allows for flexibility of the binding site.

This paper is concerned with the second step, that of designing molecules that connect a suitably chosen set of functional groups and fit well into the binding site. The proposed method has been implemented in a program called HOOK for the de novo design of ligands. HOOK begins with the three-dimensional structure of a protein binding site, and a collection of functional group sites that reflect how and where potential ligands could position functional groups to make favorable interactions with the protein. HOOK attempts to link these functional group sites together with molecular skeletons taken from a precomputed database and constructs new molecules that contain multiple functional groups in potentially favorable positions. The linkage is accomplished by fusing specified bonds in the skeleton (called "hooks") with two or more free CH_3–R bonds belonging to the functional groups. The resulting molecule has its skeleton positioned in the binding site. The fit between a given molecule and the binding site is analyzed by computing an "overlap score" based on a simplified form of attractive and repulsive van der Waals interactions. Analysis is done to rank the resulting molecules in terms of their properties. Thus the candidate ligands produced by HOOK contain functional groups that complement the chemical nature of the binding site, and have a molecular shape that complements the topology of the site.

METHODS

In this section we describe the methodology of the HOOK approach. We first indicate the type of information that is required and then describe the steps used in constructing candidate ligands. Finally, the procedural details of four HOOK calculations are described which we use to illustrate the application of the program.

Necessary Data

Three distinct classes of data are required: The atomic coordinates of the binding region of interest, a description of functional group sites in the binding region, and a database of skeleton molecules.

The *protein binding region* is the part of the protein molecule for which ligands are to be designed. HOOK requires coordinates for all atoms in this region of the protein. The region chosen should be large enough to include all charged amino acids that might effect the binding. Typically, a sphere of 20 Å radius from a central amino acid is used. The coordinates can be derived from a number of sources. The coordinates used in the examples presented here are from structures determined by X-ray crystallography. Alternatively, the coordinates for the binding region can be taken from a structure determined on the basis of NMR experiments, or model built structures can be used. When necessary, polar hydrogen atoms are added using the HBUILD[7] procedure within the CHARMM[8] program.

The *functional group sites* can be obtained from a number of sources. The first is to take a known ligand whose bound structure has been determined and to use its functional group positions. This approach can be used to test if HOOK can reproduce the original ligand from its functional group sites. The sites can also serve to find new molecules which are more rigid or fit the region better than the original ligand while still orienting the original functional groups in their known and presumably favorable positions. In cases where the structures of a number of protein–ligand complexes are known, the program could propose "hybrid" ligands which combine functional groups from various molecules. A second source of functional group sites are model building programs, which compute or assign potential positions and orientations for functional groups. MCSS is an example of such a program. It might also be possible to use strong binding sites obtained from X-ray diffraction studies of the protein crystal with a solvent composed of the functional groups of interest.[9]

For the HOOK program in its present form the functional groups must have a free CH_3 group to fuse with hooks from the skeletons in the database. A set of functional groups should be chosen to probe different properties of the protein binding region: charged groups (such as acetic acid and methyl ammonium) probe electrostatic properties, polar groups (such as methanol) probe hydrogen bonding properties, and nonpolar groups (such as ethane) probe hydrophobic properties. A feature of the MCSS approach is that each functional group site has an associated energy of interaction computed using the CHARMM potential. This energy can be used to specify a preference for the sites that will be used in a molecule, and as a selection criterion for assessing the quality of HOOK generated molecules. For functional group sites generated by other techniques, a similar interaction energy can often be assigned.

The *skeleton database* provides descriptors for the molecules that will be used to link the functional group sites. This information includes the three-dimensional atomic coordinates of the molecular structures, and the positions on the molecule of the hooks that can potentially be fused to functional groups. This database can be formed from a variety of different sources. The structures can be generated de novo, they can be taken from crystal structures such as the Cambridge Crystallographic Databank,[10] or they can come from structures generated by model-building procedures such as CONCORD.[11]

Whatever the source, two attributes of any skeleton molecule are important for the HOOK approach. The first is the availability of hooks on the molecule to which functional groups can be attached. These

Step	Figure	Notes and Parameters Used
Initialization		Identify head and tail atoms of functional group sites. Identify allowed functional group combinations where distance between atoms in pairs of functional groups should be less than 1.2 Å for head-head atoms, 1.3 Å for hydrogen bond donor/acceptor pairs and 1.7 Å for all others.
Compute internal geometry for each pair of hooks in the skeleton	Figure 3b	
Compare geometry of hooks with geometry of each pair of primary functional group sites	Figure 3b	Distance | dh-df | < 0.5 Å Angle | ah1-af1 | < 20° Angle | ah2-af2 | < 20° Dihedral | th-tf | < 30°
Overlap hooks and sites to bring skeleton into binding region	Figure 3c	RMS Overlap < 0.5 Å
Check clash of functional groups with skeleton		Pairs of atoms < 1.0 Å
Calculate contacts between skeleton and binding region		Contact distance < 4.0 Å
Compute overlap score between skeleton and binding region	Figure 3d	Overlap score for carbon-carbon interaction computed with O_{low}, O_{high} and O_{cut} of 3.2, 4.2 and 5.5 Å, respectively. Values for other atom pairs were adjusted to account for the differences in van der Waals radii of the contributing atoms, with radii for C, H, N, O and S taken as 1.6, 1.0, 1.6 and 1.8 Å, respectively. An O_{scale} value of -10 was used for all calculations. Hits were discarded if the total overlap score was below 50. The overlap score can be scaled by the number of skeleton atoms that come within 4 Å of the binding region.
Secondary search for unused skeleton hooks against all functional group sites	Figure 3e	Hook head atom - site head atom < 0.3 Å Hook tail atom - site tail atom < 0.6 Å
Additional carbon secondary search for unused skeleton hooks against all functional group sites	Figure 3f	d1 < 0.2 Å 1.24 Å < d2 < 1.84 Å 1.9 Å < d3 < 3.1 Å

Fig. 2. A flow chart showing the various stages of the HOOK algorithm, and the parameters used in the calculations presented in this paper.

are generally C–X bonds. The second is the actual conformation of the molecule, as this determines the orientation and position of the hooks on the skeleton. The current version of HOOK treats skeletons as rigid molecules. Molecules that are not rigid can be represented as replicas with different conformations.

Essentials of the HOOK Algorithm

Figure 2 is a flow chart and Figure 3 is a schematic representation of the various stages of the HOOK algorithm. The details of each of the steps in the HOOK procedure are presented in Figure 3. Many of these steps rely on parameters whose values determine the characteristics and number of candidate ligands obtained and the time required for the search. These parameters (some fixed and some user defined) and their typical (default) values are listed in Figure 2. The chosen values are based on our experience in running the program and are selected so that HOOK produces a sufficient number of acceptable molecules in a reasonable time.

The functionality map of the binding region can be considered as a collection of free CH_3–R bonds presented by each functional group site (Fig. 3a). The first step in the HOOK procedure (called the *primary search*) is to determine whether any pair of hooks from a given skeleton can be overlaid with any of the possible pairs of a selected subset of these CH_3–R bonds. This is accomplished by comparing the internal geometries of every pair of hooks with every pair of functional group sites (Fig. 3b). If all measures of internal geometry agree within certain tolerances, it is assumed that an acceptable fusion between hooks and functional group sites can be performed.

For all acceptable matches, the skeleton is oriented in the binding site by a simple least-squares refinement of matching atoms in the hooks and functional group sites (Fig. 3c). The fit of the oriented skeleton to the binding region is assessed by calculating an *overlap score* based on a simplified Lennard–Jones potential (Fig. 3d). Every possible skeleton atom–protein atom contact (excluding atoms used in the match) contributes to the score. A large positive overlap score indicates a good fit with the protein. Overlap scores below a given value result in the rejection of a match.

All oriented skeletons that pass the overlap score test are used in a *secondary search* to determine if any additional functional group sites can be attached to the skeleton in its current orientation.

Two types of additional linkages are considered (Fig. 3e): the fusion of unused hooks and functional group sites, or the linkage of hooks and functional groups through a new bond. Such matches can be made if certain geometric criteria are met by the hooks and functional group sites (Fig. 3f).

Following the secondary search, potential candidate ligands (HOOK hits) are generated and stored in an output file along with most of the numerical information used in generating the molecule. This output file is analyzed by the program TABLE, which allows interactive manipulation of the data from the HOOK search. Certain spreadsheet-like functions are provided for analyzing and ordering the hits. The analyses can be made on the basis of stored information from the HOOK calculations, such as the overlap score, or by means of criteria generated with the TABLE program. Examples of the types of data operations are given with the results. TABLE also can be used as an interface to other calculations, e.g., it can construct molecules for use in programs, such as QUANTA (Molecular Simulations, Inc.), which provide an environment for viewing the molecules, or for further analysis with programs, such as CHARMM.[8]

Systems Examined

Binding regions

Two different binding regions are studied in the examples presented in this paper. The first is from the hemagglutinin (HA) molecule of the influenza A virus. One function of HA is to bind to cell surface sugars terminated with sialic acid. The structure of a sialic acid–HA complex has been determined by x-ray crystallography,[12,13] and this is the region we study here. The second binding region is from chloramphenicol acetyltransferase (CAT), an enzyme that catalyzes the acetylation of chloramphenicol by acetyl-CoA.[14] The coordinates used in this study were taken from the Brookhaven Protein Databank[15] (code 4HMG for HA—3.0 Å resolution, and 1CLA for CAT—1.7 Å resolution). In both cases, the binding region of interest is localized in a well-defined part of the structure, and a sphere of approximately 20 Å was used to define the region of the proteins used in the analysis. All other atoms were deleted. For HA the sphere was centered around Trp-153 at the base of the sialic binding region. For CAT, the sphere was centered around the catalytic His-84 in the chloramphenicol binding region, which is at the trimer interface. Polar hydrogens were added to the coordinates that make up the binding regions. No further minimization was done on either coordinate set, and the protein was kept rigid during all the studies. The final binding region coordinate sets (including polar hydrogens) contained 1500 atoms for HA and 2290 atoms for CAT. All of the binding region was used for the MCSS calculations on both molecules. For HA, a separate subset of 393 atoms was identified for use in computing overlap scores. It consisted of all the amino acid residues for which an atom is solvent accessible within 15 Å of the central Trp-153 of the binding region. For CAT, all 2290 atoms were used, so the overlap score calculation time is much longer; the approach employed for HA gives satisfactory results at a much lower cost.

Functional group sites

The functional group sites considered in these examples arose from two sources. The first source is the positions of functional groups in the known crystal structure of the complex of sialic acid and HA. The sialic acid molecule in this structure is a derivative of *N*-acetyl neuraminic acid, and seven functional group sites can be identified as shown in Figure 1b. They include one acetic acid, one *N*-methyl acetamide, and five methanol groups. *N*-Methyl acetamide contains two CH_3–R bonds that could fuse with a skeleton hook. However, in the case of sialic acid, the methyl attached to the carbonyl group is essentially buried in the HA binding region, so only the methyl attached to the amide group was designated as a potential functional group site. The second source of functional group sites is obtained by use of the MCSS program. The possible interactions of four different functional groups with the HA and CAT binding regions were studied with MCSS. These functional groups are acetic acid, methanol, *N*-methyl acetamide, and methyl ammonium, as shown in Figure 1a. As mentioned above, two input functional group sites can be generated from each *N*-methyl acetamide in the functionality map from the methyls attached to the amide and carbonyl, respectively.

The MCSS-generated functional group sites for HA and CAT that were used in these examples are summarized in Table I and discussed in more detail in the results section of the paper. Division into primary and secondary functional groups sites was based on observed clusters in the interaction energies of the sites; Table I indicates the energy ranges chosen for each study.

Skeleton database

The program SKELETON has been developed to construct skeletons and define hooks from molecules in any database containing three-dimensional structures of small molecules. This process can be as simple as describing each molecule in skeleton database format (coordinates of atoms with hooks defined) or it may involve identifying rigid substructures within each molecule and describing each of these substructures in the appropriate format.

Two different databases were used to provide the skeleton molecules used in the examples presented here. The first database is composed of small hydro-

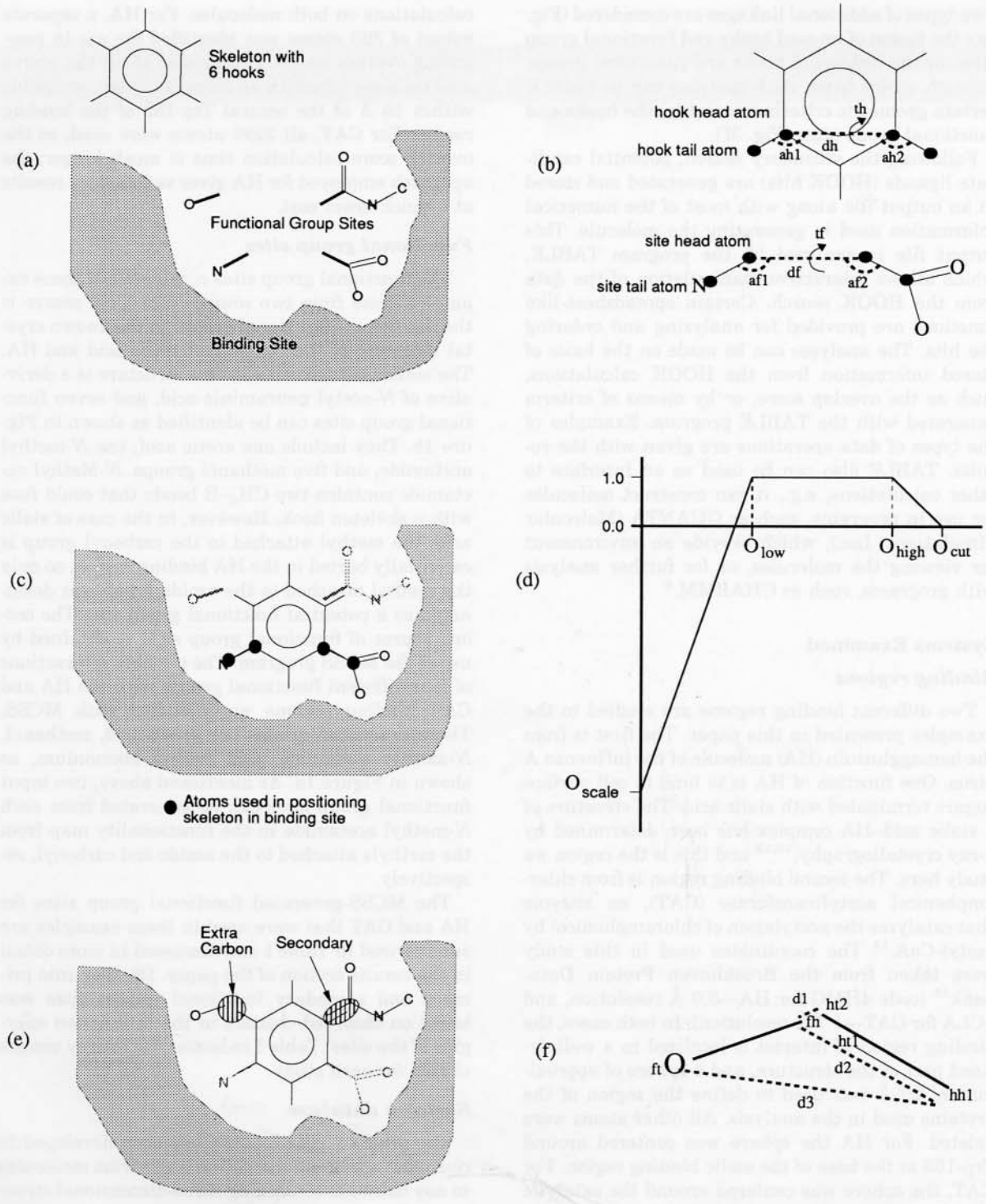

Skeleton with
6 hooks
(a)
Functional Group Sites
Binding Site
(b)
hook head atom
hook tail atom
th
dh
ah1
ah2
site head atom
site tail atom N
tf
df
af1
af2
O
(c)
Atoms used in positioning
skeleton in binding site
(d)
1.0
0.0
O low
O high
O cut
O scale
(e)
Extra
Carbon
Secondary
(f)
d1
ht2
fh
ht1
d2
ft O
d3
hh1

Fig. 3.

Fig. 3. A schematic of the HOOK algorithm. (**a**) The functionality map is considered to be a collection of free CH_3–R bonds capable of being fused to a skeleton. Note that some functional groups (such as *N*-methyl acetamide) have more than one such CH_3–R bond; each is considered separately in HOOK. (**b**) Possible matches of pairs of hooks and pairs of functional group sites are considered by comparing internal geometries determined by four atoms from each pair. Each hook is described by a hook head atom (usually carbon) and a hook tail atom (usually hydrogen). Each functional group site is described by a site head atom (the carbon of the CH_3 group) and a site tail atom (the attached atom of the R group). Four measures of internal geometry are calculated for each set of four atoms: the distance between head atoms, the two angles determined by one tail atom and both head atoms, and the dihedral determined by all four atoms. A match is accepted only if all four measures agree within certain user defined tolerances (see values in Fig. 2). Since few skeletons are perfectly symmetric, the two possible ways of matching pairs of hooks and pairs of sites are treated. (**c**) After a suitable match is found, the skeleton is positioned in the binding region using a transformation determined by the least squares superposition of the atoms used in the internal geometry calculations. The hook tail atoms used in this calculation are actually dummy atoms whose positions are calculated so that the length between the hook head atom and tail are the same as the bond length between the site head atom and tail atom. If the rms difference of the two sets of coordinates exceeds a user defined tolerance, the orientation is rejected. Since only two atoms from each functional group site are considered in the overlap, it is possible that atoms in the oriented skeleton will fall too close to some of the remaining atoms in the functional group. If any such contacts occur (as determined by a user defined close contact tolerance) the orientation is rejected. (**d**) The overlap score is based on a simplified Lennard–Jones potential. The interaction between every atom of the skeleton (excluding atoms involved in the ligated hooks) and every atom of the binding region makes a contribution to the overlap score of the oriented skeleton. The contribution of each pair is calculated from the interatomic distance, and its relation to four parameters (O_{cut}, O_{high}, O_{low}, and O_{scale}). If the distance is greater than O_{cut}, the contribution is zero. For distances between O_{cut} and O_{high}, the contribution increases linearly to one. It remains fixed at one for distances between O_{high} and O_{low}, and then falls off linearly to $-O_{scale}$ as the distance approaches zero. The relative values of O_{cut}, O_{high}, O_{low}, and O_{scale} can be adjusted for different runs to vary the general behavior of the function. The actual values of O_{cut}, O_{high}, and O_{low} are adjusted for each atom pair in accord with the van der Waals radii of the atoms involved (see Fig. 2). Since the overlap score is biased in favor of larger skeletons, HOOK allows for the score to be scaled by either the number of atoms in the skeleton or the number of skeleton atoms within a defined distance of the binding region. (**e**) Once a skeleton is positioned in the binding region and it has been determined that the fit of the oriented skeleton to the binding region (as determined by the overlap score) is satisfactory, a search is made for additional functional group sites that can be attached to the remaining hooks of the skeleton. Two types of extra linkage are possible. The first (labeled secondary) is a fusion of a hook and an additional functional group CH_3–R bond. The criteria for adding a functional group site by this method are that the distance between the hook head atom and the site tail atom, and the hook tail atom and the site head atom must both lie within given tolerances (again, dummy atoms are used to normalize bond lengths). The second type of linkage (labeled extra carbon) involves the creation of a new carbon–carbon bond between the hook head atom and the site head atom. Such a linkage is deemed possible if the geometric criteria for the creation of such a new bond [as oulined in (**f**) below] are satisfied within given tolerances. As in the primary search, if the addition of the new functional group would result in the overlap of additional atoms from the functional group site and the skeleton, it is rejected. It is possible that more than one possible additional group can be identified for a given hook. As currently implemented, HOOK requires that only one such group be accepted, and a number of possible methods for making the choice are provided. (**f**) The geometric criteria for adding a functional group through a new carbon–carbon bond. There are three distances (d1, d2, d3) that have optimal values for a C–C–R linkage. If the observed values are close to these optimal values, it is possible to add the functional group to the skeleton through such a bond.

carbon skeletons constructed from a set of saturated mono-, bi-, and tricyclic hydrocarbons with ring sizes of 5 and 6 carbons. All possible unsaturated derivatives of these molecules were generated (M. Eisen, unpublished program). The geometries of these unsaturated molecules were regularized (i.e., correct bond length and bond angles and reasonable dihedral angles were obtained) by minimizing them with the program CHARMM. The central pyranose ring skeleton and attached glycerol side chain of the sialic acid molecule found in the crystal structure of hemagglutinin was added to database A which has a total of 980 skeletons, 22,957 atoms, and 12,303 hooks; there are 9 monocyclic, 361 bicyclic, and 605 tricyclic hydrocarbons in the set.

The second database, database B, was constructed from the Cambridge Crystallographic Databank[11] by selecting a subset of molecules which contains at least two of the four functional groups used in the MCSS calculations. This choice was made to ensure that the database was small and included only a selection of organic molecules which could link the appropriate types of functionality. A total of 6012 molecules was obtained. Skeleton substructures were identified from these molecules using the SKELETON program. In this analysis, single bonds between atoms which are not in a ring were identified for each molecule. The bonds were removed, any nonbonded atoms deleted, and hydrogen atoms added to complete valency before storing the resulting molecular fragments. The resulting list of fragments was sorted and duplicates were eliminated. Database B has 1875 unique skeletons containing between 6 and 98 atoms each and a total of 36,937 atoms and 11,125 hooks.

In both database A and B, every C–H bond was identified as a hook with the potential to link to any functional group. Thus, there are relatively fewer hooks in database B due to the presence of polar atoms.

For all the HOOK runs done here, it was required that at least one functional group was added to the molecule in addition to the two primary groups. The other parameters were as given in Figure 2.

RESULTS

Here we present the results of a series of calculations using HOOK to construct molecules in the active sites of HA and CAT, using the positions of functional groups either from known crystal structures or from MCSS calculations. The results have been analyzed to demonstrate the types of molecules that are generated and to characterize the distinctive features of the program.

Sialic Acid Functional Groups in HA (Database A)

The simplest test of the HOOK strategy is to use the position and orientation of functional groups

TABLE I. Summary of MCSS Results for HA and CAT

Hemagglutinin Center of binding region—center of Trp-70 ring Radius of binding region—12 Å			
Functional group	Number of sites (energy cutoff)*	cpu time (min)	Number of primary sites (energy cutoff)*
Acetic acid	12 (−15)	320	12 (−50)
N-Methyl acetamide	49 (−8)	412	6 (−21)
Methanol	31 (−8)	236	6 (−18)
Methyl ammonium	14 (−15)	330	5 (−50)
Chloramphenicol acetyltransferase Center of binding region—NE^2 of His-84 Radius of binding region—10 Å			
Functional group	Number of sites (energy cutoff)*	cpu time (min)	Number of primary sites (energy cutoff)*
Acetic acid	14 (−15)	337	4 (−40)
N-Methyl acetamide	29 (−8)	427	17 (−21)
Methanol	34 (−8)	231	2 (−18)
Methyl ammonium	12 (−15)	254	4 (−40)

*Interaction energy between functional group site and protein binding region as calculated by MCSS in kcal mol^{-1}.

from a known protein–ligand complex as the input functional group sites. This test serves both to validate the implementation of the HOOK approach by regenerating known molecules, and to offer a simple framework to understand how HOOK can suggest novel architectures for linking known functional group sites. The sialic acid–HA complex is a useful model for such a test, as seven distinct functional group sites can be derived from the crystal structure of bound sialic acid (Fig. 1b).

In this first example, HOOK searched database A (which contains cyclic hydrocarbon molecules and the pyranose core of sialic acid) for skeletons that could link the sialic acid functional group sites in the context of the HA binding region. Table II summarizes the results of this and the other HOOK runs presented here. From the database of 980 skeletons, 9222 hits were found which had overlap scores greater than 50 and had at least three functional group sites attached. Using the TABLE program, these hits were sorted first on the number of extra functional groups they contained, with a secondary sort on overlap score. Twenty of the top 24 hits were made by the sialic acid skeleton fusing with all 7 of the functional group sites. They are based on the 20 unique pairs of functional group sites, excluding the pair consisting of the acetic acid and methanol 2.

The primary search of the HOOK program can explicitly ignore a combination of two functional group sites that are close together. This facility was used here, otherwise every CH_2 group in every skeleton would provide two hooks that matched the geometries of the acetic acid and methanol 2 functional group sites. However, in the secondary search, this restriction was lifted, so that all of the primary hits that incorporated either the acetic acid or methanol 2 could find the other site.

The present HA/sialic acid example produced a large number of hits because the functional group sites used are in a reasonable position and orientation for linkage by the many molecules with structures similar to the sialic acid pyranose cores in database A.

Figure 4a shows some other representative hits together with an indication of the sialic acid functional group sites they use, their overlap score, and their position in the sorted list of HOOK hits. For reference, the sialic acid overlap score in this binding region is 124. Molecule 1 is the highest scoring hit connecting all 7 functional group sites. It is closely related to sialic acid, in that the carbocyclic homologue of the pyranose ring is present along with two of the carbons of the glycerol side chain. The difference is that the side chain is stabilized by having two of its carbons as members of a second carbocyclic ring. This molecule spans the binding region to the same extent as sialic acid by linking to methanol 5 through an additional CH_2 group, introduced by the HOOK program during the secondary search for extra functional group sites. HOOK performed the primary search for each pair of hooks; for four of these pairs, the overlap score is higher than that for sialic acid. The other combinations are found lower down the HOOK list. They vary slightly in overlap score, because the skeleton is oriented in the binding region on the basis of the overlap of a given pair of hooks with a particular pair of functional group sites. Different pairs of hooks give slightly different orientations and small changes in orientation can modify the overlap score. Introduc-

TABLE II. Summary of HOOK Runs for HA and CAT

	Calculation number			
	1 HA + sialic acid functional sites	2 HA + MCSS sites	3 HA + MCSS sites	4 CAT + MCSS sites
Database	A	A	B	B
Number of skeletons	980	980	1875	1875
Number of skeleton hooks	12303	12303	11157	11157
Number of functional sites (Number of primary sites) (see Table I for details)	7 (7)	106 (29)	106 (29)	118 (44)
Time (sec) for HOOK run (single processor on SGI 380)	2142	2166	1788	19620
Number of hits	9222	2295	460	655
Highest overlap score	211	226	330	388

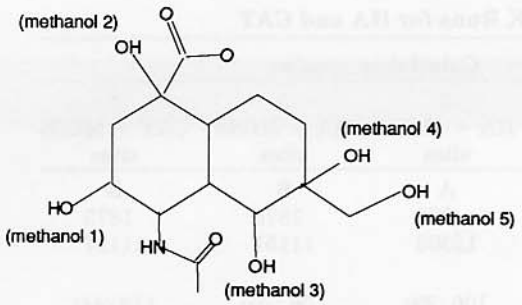
ing a double bond in the molecule as in molecule 2 links together the same functional group sites, but gives a slightly reduced overlap score. Inspection of the position of molecule 2 in the binding region shows that the small change in conformation of the bicyclic ring changes the complementarity of the molecule with the HA binding region. Molecule 3 has the highest overlap score for a molecule linking 4 functional group sites. It is similar to molecules 1 and 2, but contains three double bonds. The resulting change in conformation of the skeleton prohibits linkage to methanol 4. Finally, molecule 4 is the highest scoring molecule containing a five-membered ring. The pattern of linkages to functional group sites is quite different in molecule 4 compared to that for molecules 1–3, e.g., hydroxyl 4 and 5 are linked to the same carbon of the main skeleton in molecule 1 while they are linked to different carbons in molecule 4. This is illustrated in Figure 6, which shows molecules 1 and 4 in the binding region of hemagglutinin.

This selection of molecules demonstrates the variety of architectures which can link even a small set of functional group sites. Using HOOK on such a collection of sites may prove to be a useful means of identifying novel architectures for known sets of functional group sites, or perhaps offer a better lead compound for improvement. It is also a potential means for using information from multiple ligands. The collection of functional group sites derived from all the ligands could form the input for a HOOK calculation to produce new molecules.

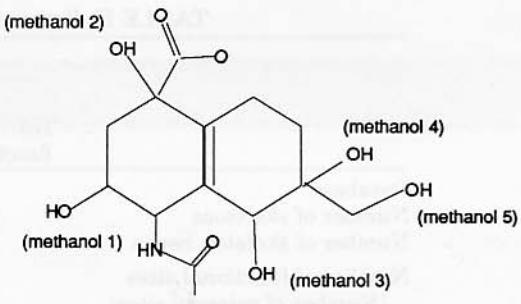
In this HOOK run, the overlap score of each molecule was an important aspect of how these molecules were prioritized. To demonstrate that this score successfully identifies acceptable binding positions, we analyzed the variation in the overlap score of the fit of sialic acid to the HA binding region as the sialic acid was moved from its known crystal position. Figure 5a shows the sialic acid in the HA binding region in the crystal structure.[12,13] Figure 5b is a contour plot showing how the overlap score varies as the sialic acid is moved in the indicated x–y plane of Figure 5a. It can be seen that the position with the highest overlap score is very close to the crystal position, and that the score drops rapidly as the ligand moves away from that position. The extent to which close contacts are penalized can be varied by changing the penalty value O_{scale} (see Fig. 3d). In the examples presented here, an O_{scale} value of -10 was used.

To assess the effect of changing this value between -2 and -50, a series of HOOK calculations were performed for the sialic acid functional group sites and database A. Table III shows the results. It gives the total number of hits with an overlap score greater than 50 (T) and the position of sialic acid in this list ordered on overlap score (P). An O_{scale} value of -10 gives the best ratio T/P for the relative position of the sialic acid in the list. Decreasing this value to -50 excessively penalizes good contacts between the ligand molecule and the protein, resulting in the overlap score being highest for molecules that are not making close contacts with the binding region. Increasing the value to -2 produces a significant number of molecules that by inspection have unsatisfactory overlaps with the binding region.

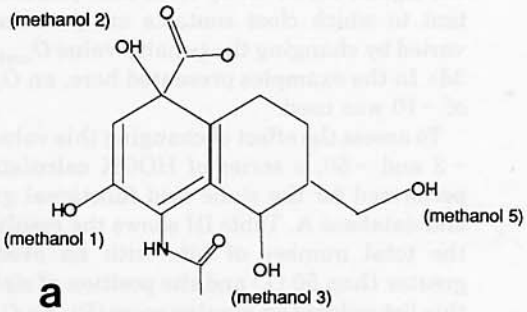
Table IVa summarizes the number of HOOK hits found which link different numbers of sialic acid functional groups sites in hemagglutinin. A relatively small number of molecules can link all seven of the sites; 20 out of 30 of these involve the sialic acid skeleton. It is clear that at least some of the 10 others are good candidates for investigation as possible ligands. Table V summarizes the distribution of overlap score for the HOOK hits. The hits with the highest overlap scores, as well as the lowest, are from skeleton molecules that link together just three of the functional group sites. With appropriate parameters HOOK clearly finds a wide range of compounds that have many functional groups and make good interaction with the binding region.



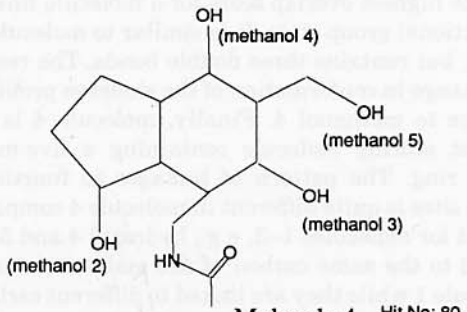
Molecule 1 Hit No: 1–4 Score: 129



Molecule 2 Hit No: 25–26 Score: 122



Molecule 3 Hit No: 31 Score: 134

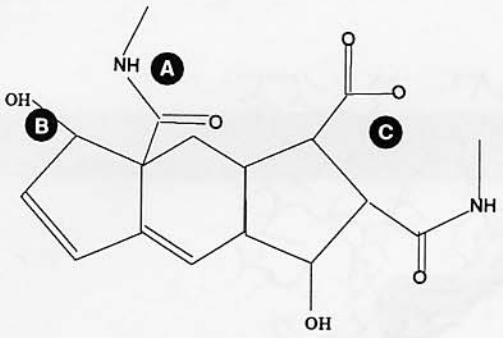


Molecule 4 Hit No: 80 Score: 131

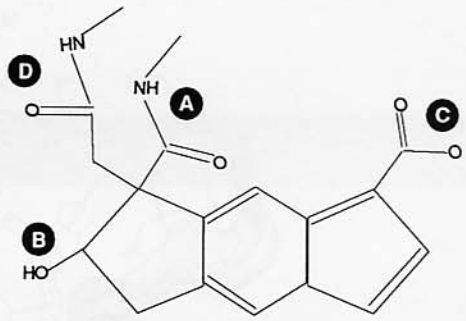
Fig. 4. Molecules constructed by HOOK from database A skeletons for the hemagglutinin active site. (**a**) Four of the molecules generated by HOOK which link together some of the seven functional sites identified in the crystal structure of hemagglutinin–sialic acid. (**b**) Seven of the molecules generated by HOOK which link together some of the functional sites computed by MCSS. The symbols A–E identify where the same functional group sites are used in different molecules. The position in the hit list (as processed by the TABLE program and described in the text), the overlap score, and the overlap score per skeleton atom are given.

MCSS Functional Groups in HA

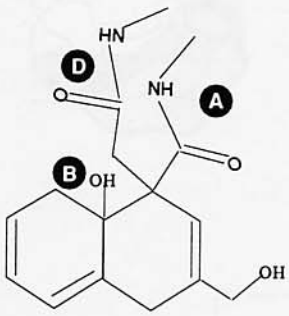
For the next test of the approach we used HOOK to generate molecules that link functional group sites identified by MCSS. MCSS functionality maps of the sialic acid binding region of HA were generated for the four functional groups acetate, methyl ammonium, methanol, and *N*-methyl acetamide. The times required for the MCSS calculations and a summary of the results are presented in Table I. The time varied with the nature of the functional group, and ranged from 4 hr for methanol to 7 hr for *N*-methyl acetamide when run on a single processor Silicon Graphics 380 series computer. The interaction energy between each functional group site and the protein binding region is computed during the MCSS procedure. These were used to partition the minima into primary and secondary functional group sites. Acetic acid and methyl ammonium can make strong electrostatic interactions with the protein binding region, so that sites with an interaction energy less than -50 and less than -15 kcal mol^{-1} were chosen as primary and secondary sites, respectively. Methanol and *N*-methyl acetamide can make a variety of polar and hydrogen bonding interactions with the protein binding region, and functional groups with an interaction energy less than -18 and -21 kcal mol^{-1}, respectively were chosen as primary sites. Secondary sites for methanol and *N*-



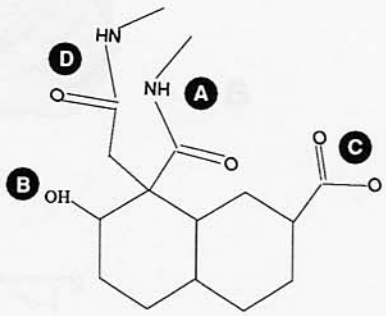
Molecule 5 Hit No: 1–2
Score: 93
Score/Atom: 3.9



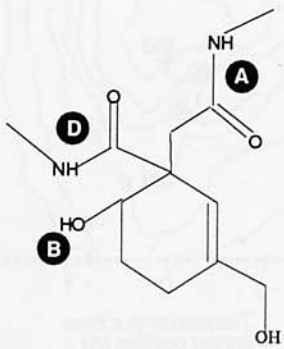
Molecule 6 Hit No: 3
Score: 183
Score/Atom: 7.6



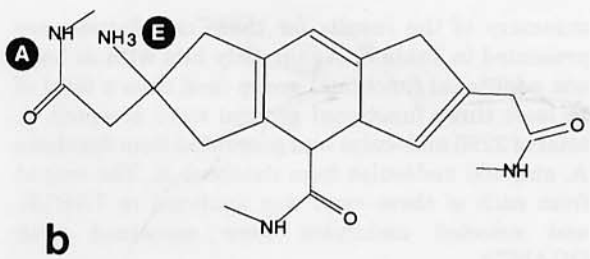
Molecule 7 Hit No: 9
Score: 170
Score/Atom: 7.7



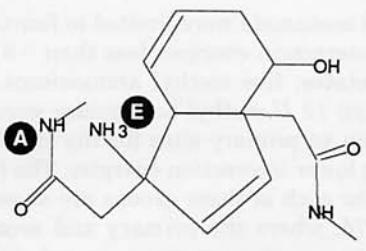
Molecule 8 Hit No: 12
Score: 161
Score/Atom: 5.7



Molecule 9 Score: 136
Score/Atom: 8.5



Molecule 10 Score: 145
Score/Atom: 5.6



Molecule 11 Score: 140
Score/Atom: 6.7

Figure 4b.

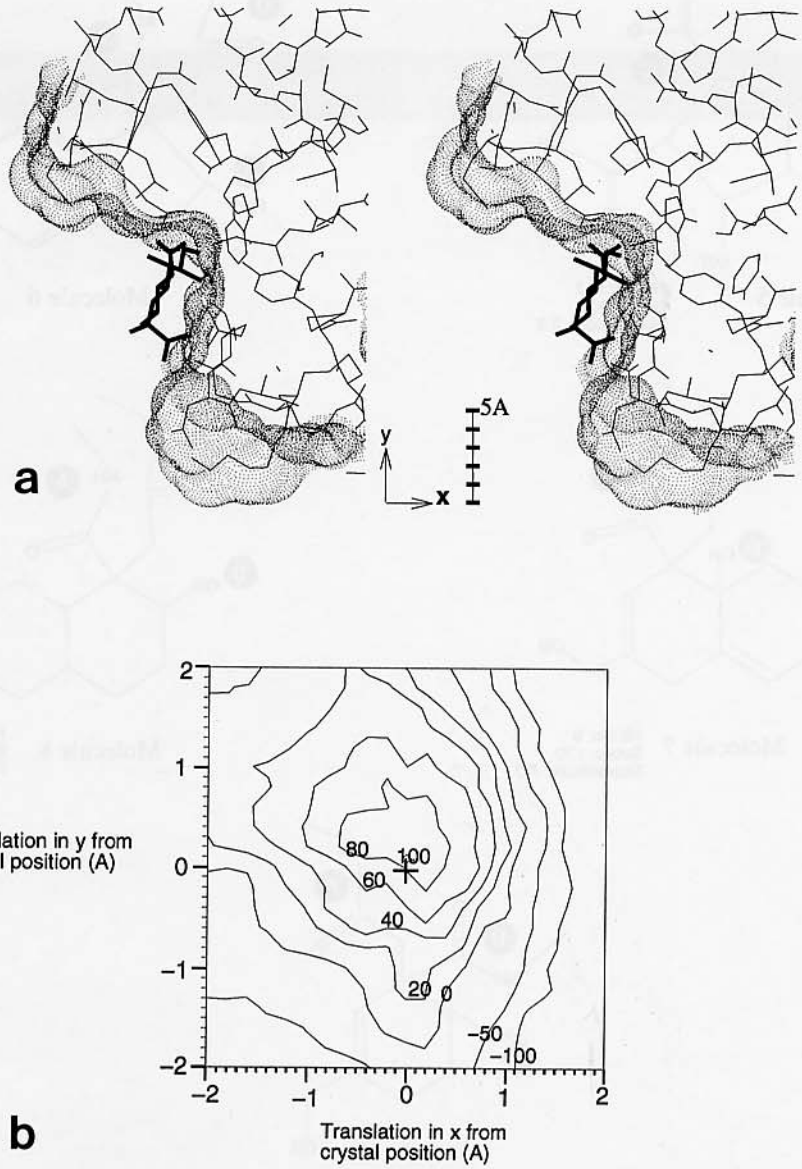


Fig. 5. An assessment of the overlap score for a protein–ligand complex of known structure. (**a**) Stereo representation of the structure of sialic acid bound to the haemagglutinin active site as seen in the crystallographic analysis of Weis et al.[13] Sialic acid is shown in bold lines. The line represents 5 Å. (**b**) The overlap score for sialic acid as it is moved within the HA binding region vertically and horizontally in the view indicated in (**a**).

methyl acetamide were limited to functional groups with interaction energies less than -8 kcal mol^{-1}. Six acetates, five methyl ammoniums, six methanols, and 12 *N*-methyl acetamides were selected in this way as primary sites for the HOOK search as having lower interaction energies. The functionality maps for each of these groups are shown in Figure 7a to 7d, where the primary and secondary functional group sites are shown in red and yellow, respectively.

HOOK was used to search both database A and database B for skeletons that could link the MCSS functional group sites in the HA binding region. A summary of the results for these calculations are presented in Table II. Again, only hits with at least one additional functional group (and thus a total of at least three functional groups) were accepted. A total of 2295 molecules was generated from database A, and 460 molecules from database B. The output from each of these runs was analyzed in TABLE, and selected molecules were examined with QUANTA.

Database A

Figure 4b shows a selection of molecules constructed by HOOK from the skeletons of database A,

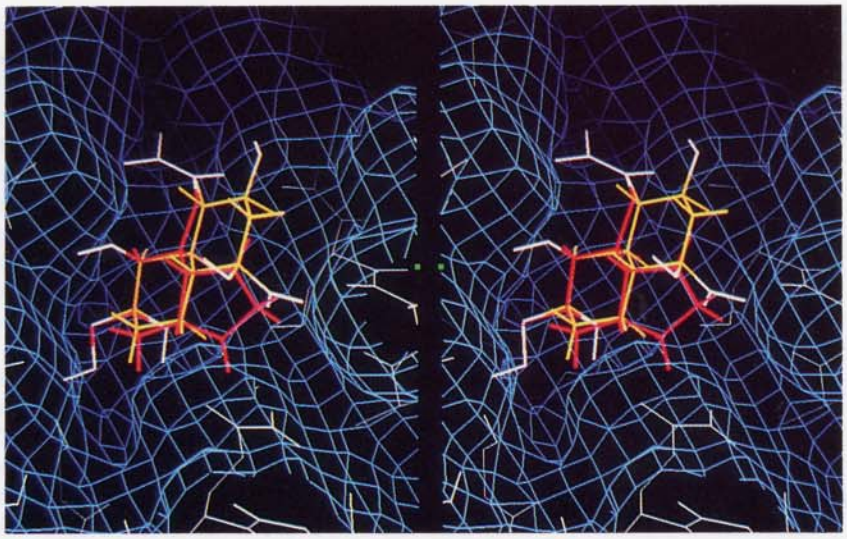


Fig. 6. Molecules 1 and 4 of Figure 4a as built by HOOK in the HA binding region. The molecular surface of HA computed on a 1-Å grid at a radius of 0.8 Å above the van der Waals surface of the molecule is shown in light blue, HA structure in thin light green. The skeleton of molecule 1 is red, the skeleton of molecule 4 yellow, with the functional group sites shown in white.

TABLE III. Variation of Overlap Score With O_{scale}*

O_{scale}†	Overlap score of SA	Highest score	Position of SA in list by overlap score (P)	Total number of hits with score >50 (T)	T/P
−2	139	367	6907	10929	1.58
−5	133	260	5719	10907	1.91
−10	129	211	1835	9222	5.03
−20	106	176	1277	5042	3.95
−50	51	92	2673	2763	1.03

*Variation of the overlap score for sialic acid (SA) in its known position in the hemagglutinin binding region as the O_{scale} parameter is changed. The results are for HOOK runs for the sialic acid functional groups in the HA binding region with database A. All other HOOK parameters as given in Figure 2.
†See Figure 3d for definition.

and the functional group sites generated by MCSS. The symbols A to E correspond to identical functional group sites in different molecules; additional functional group sites that are unique for given molecules are present in most cases (e.g., an OH and NMA on molecule 5). These molecules were selected using the TABLE program in the following way. The hits were first ordered on the number of functional group sites followed by the overlap score (as in the sialic acid functional site example discussed above). Only molecule 5 links five functional group sites. However, the overlap score for this molecule is quite low, mainly because of repulsive interactions between atoms in the protein binding region and atoms in the cyclopentane ring that contains hydroxyl B. Molecules 6 and 8 are selections from the molecules that combine with four functional group sites. Molecule 6 has the best overlap score for molecules connecting four sites. The comparison of molecules 7 and 8 shows that different saturation patterns in the bicyclic hydrocarbon skeletons can yield subtly different orientations of the C–H bond hooks. Although the functional group sites A and D appear to be close, they are actually pointing away from each other. This results in a variation in the pattern of linkages of the functional group sites and thus the molecules have different positions in the binding region. This is reflected in the different overlap scores obtained. An interesting comparison involves molecules 6 and 8 that have different skeletons but the same functional group sites. Ordering the hits on the number of functional group sites first followed by use of the overlap score scaled by the number of atoms in the skeleton, gave (after molecule 5) the monocyclic molecule 9 as the best hit.

Table IVb summarizes the number of HOOK hits

TABLE IV. Profile of HOOK Hits

Number of sites	Number of hits	Overlap score range
(a) Sialic Acid Functional Group Sites in HA With Database A		
7	30	129–124
6	16	133–85
5	163	158–58
4	1587	197–50
3	7426	211–50
(b) MCSS Functional Group Sites in HA With Database A		
5	2	93
4	86	183–52
3	2207	226–50
(c) MCSS Functional Group Sites in HA With Database B		
5	3	220–197
4	25	228–53
3	432	330–51
(d) MCSS Functional Group Sites in CAT With Database B		
5	6	340–61
4	71	388–54
3	578	380–50

TABLE V. Profile of Overlap Scores

System	> 200	> 150	> 100	> 50
Sialic acid HA Database A	8	538	3856	4820
MCSS HA Database A	11	374	1299	611
MCSS HA Database B	17	86	174	183
MCSS CAT Database B	98	97	166	294

which link different numbers of functional group sites. In the example with sialic acid functional groups (Table IVa), there were a significant number of molecules which linked between 5 and 7 sites because the groups were in the same orientation as many of the skeleton hooks. The orientation of the MCSS derived functional group sites is more varied and does not include all of those from sialic acid. This was noted in the original MCSS study[4] where it was suggested that the sialic acid functional group sites are not optimal, in accord with the fact that sialic acid is a poor ligand. Table V shows that if a skeleton can link several functional group sites, the overlap score is generally favorable.

Functional group sites A and/or B (see Fig. 4b) are present in many of the HOOK hits. Of the 2295 molecules found by HOOK, some 1280 contain at least one of these groups and 572 contain both, suggesting that the particular relative orientation of these two functional group sites is such as to satisfy the internal geometry of many skeleton architectures. Generally, one of the groups is linked directly to the skeleton and the other is linked through an additional CH_2 group.

One of the features of the MCSS approach is the calculation of an energy of interaction between the functional group and the protein. The list of HOOK hits from this MCSS run was also sorted by the TABLE program, firstly by the total energy of the functional group sites linked by the skeleton, followed by the overlap score divided by the number of atoms in the skeleton. This gave molecules 10 and 11 as the best hits; their MCSS interaction energy is −137 and −135 kcal/mol, respectively. These molecules contain only three functional group sites, but each of them has a strong interaction energy with the binding region. The dominant contribution comes from ammonium group E in both molecules. An alternative way of sorting molecules would be in terms of the calculated functional group interaction energies, relative to the functional group solvation energies,[4] to "renormalize" the large differences in the energy ranges involved.

Database B

In the second run with the MCSS functional group sites in HA, database B was used with the same set of MCSS minima. In the HOOK calculations, only 460 hits were found which satisfy the HOOK criteria used here (three or more linked functional groups and an overlap score greater than 50). In database A, most of the skeletons are compact. For database B many of the skeleton molecules have rather complex geometry and stereochemistry which increases the likelihood of making unfavorable overlap with the binding region when linked to the functional group sites. The database, therefore, provides an example of how the HOOK approach can suggest novel architectures for satisfying the restraints of a binding region.

The TABLE program was used to assess these 460 structures which were ordered on the total interaction energy of the attached functional group sites, followed by the overlap score per atom in the skeleton molecule. The top 10 molecules contained a variety of different architectures. Figure 8a shows the best hit which is a cyclodextrin derivative that links together an acetic acid, a methyl ammonium, and an *N*-methyl acetamide with a combined MCSS interaction energy of −183 kcal/mol^{-1}. The cyclodextrin

ring fits into the base of the binding region. The two attached benzene groups are positioned to fill other parts of the region, with one linked to an acetate group on the outer edge of the binding region. Although the fit of the molecule to the binding region is not perfect, there are significant areas of good steric overlap. This shows how a complex molecule can fit into the site. Since there are a number of oxygens in the skeleton, a more detailed analysis of how they interact with the site would be required to evaluate it as a candidate ligand.

The number of molecules that have a given number of functional groups linked to a single skeleton found here (Table IVc) is similar to that obtained with database A (Table IVb). However, the profile of overlap scores is different (Table V). The larger molecules in database B tend to be more difficult (lower overlap score) to accommodate in the HA binding region. Figure 9a shows two representative hits (molecules 12 and 13), chosen to highlight how skeleton molecules of quite different shapes can link together functional group sites. A number of hits contain the basic skeleton of molecule 12 which consists of several linked saturated C6 rings.

MCSS Functional Groups in CAT (Database B)

One of the limitations of the HA binding region as an example for testing the HOOK approach is that it is very open. It is of interest to use HOOK and the overlap score criterion to identify molecules that can bind to a geometrically more restricted region. Such an example is supplied by the chloramphenicol binding region in CAT. Chloramphenicol is buried deep in the interface between two protein molecules in the active trimer. Compared to the sialic acid binding region of HA, such a buried active site provides a challenge for HOOK, in which the screening of steric interaction via the overlap score is essential in finding possible ligands. In HA a high overlap score was important to show that a molecule made good contacts with the site; for CAT, the dominant need is to avoid bad contacts.

The timing of the MCSS calculations and a summary of results are given in Table I for the four functional groups, acetate, methanol, methyl ammonium, and *N*-methyl acetamide in the binding region of CAT. Using the same criteria as in the MCSS calculation on HA, 14 acetic acid, 29 *N*-methyl acetamide, 34 methanol, and 12 methyl ammonium functional group sites were chosen for use in the HOOK program. The distribution between groups and number of sites is similar to that for HA. As the CAT binding region contains fewer charged amino acids, the interaction energies for charged functional group sites were somewhat lower than in the HA calculations. For this reason values of the interaction energy equal to -40, -21, -18, and -40 kcal mol^{-1}, respectively, were used to select 4 acetic acids, 17 *N*-methyl acetamide, 2 methanol, and 4 methyl ammonium sites as primary sites.

The results of the HOOK calculations with database B are summarized in Table II. The results were again analyzed in TABLE by sorting on the basis of the total interaction energy of the functional group sites attached to the skeleton, followed by the overlap score per skeleton atom in contact with the protein. Figure 8b shows the best molecule found by using these criteria. This is formed from a 2,4-diphenyl-bicyclo[3,3,1]nonane skeleton. This links together an *N*-methyl acetamide and a methyl ammonium on each penyl ring, but with different patterns of substitution. This molecule fits the constraints of the CAT binding region particularly well, the overlap score is 215 and the total MCSS interaction energy is -159 kcl mol^{-1}. Figure 9b shows the next two best hits. Molecule 14 has a good total energy of interaction between the linked functional groups and the binding region, but has a low overlap score. Molecule 15 has the same skeleton as molecule 12 for HA. This long molecule is threaded through the long, thin chloramphenicol and acetyl-CoA channel in CAT, demonstrating how well the HOOK approach screens for molecules that optimize their steric interaction with the binding region surface.

The profile of the number of functional group sites linked per skeleton (Table IVd) is similar to that of HA, but the distribution of overlap scores (Table V) is markedly different, with a higher proportion of both high and low scores. This is a consequence of the enclosed nature of the CAT binding region.

DISCUSSION

The HOOK program has been developed to generate molecules that satisfy certain conditions for binding to a region of interest; in the present case it is the receptor or active site of a protein of known structure. The candidate ligand molecules are designed to make favorable functional group interactions with the protein and to have a shape that complements the site. An important feature of the HOOK approach is that it efficiently identifies molecular skeletons that link together a number of functional groups selected from a large number of discrete sites. The program was developed to exploit the results of the MCSS methodology which provides the positions and orientations of the functional group sites.

In the discussion we first describe some of the elements of the program that are essential for the efficient generation of candidate molecules. We then consider possible extensions of HOOK. Finally we compare the HOOK program with other methods that have been proposed, most of them rather recently, for the design of ligands that bind to the sites of proteins of known structure.

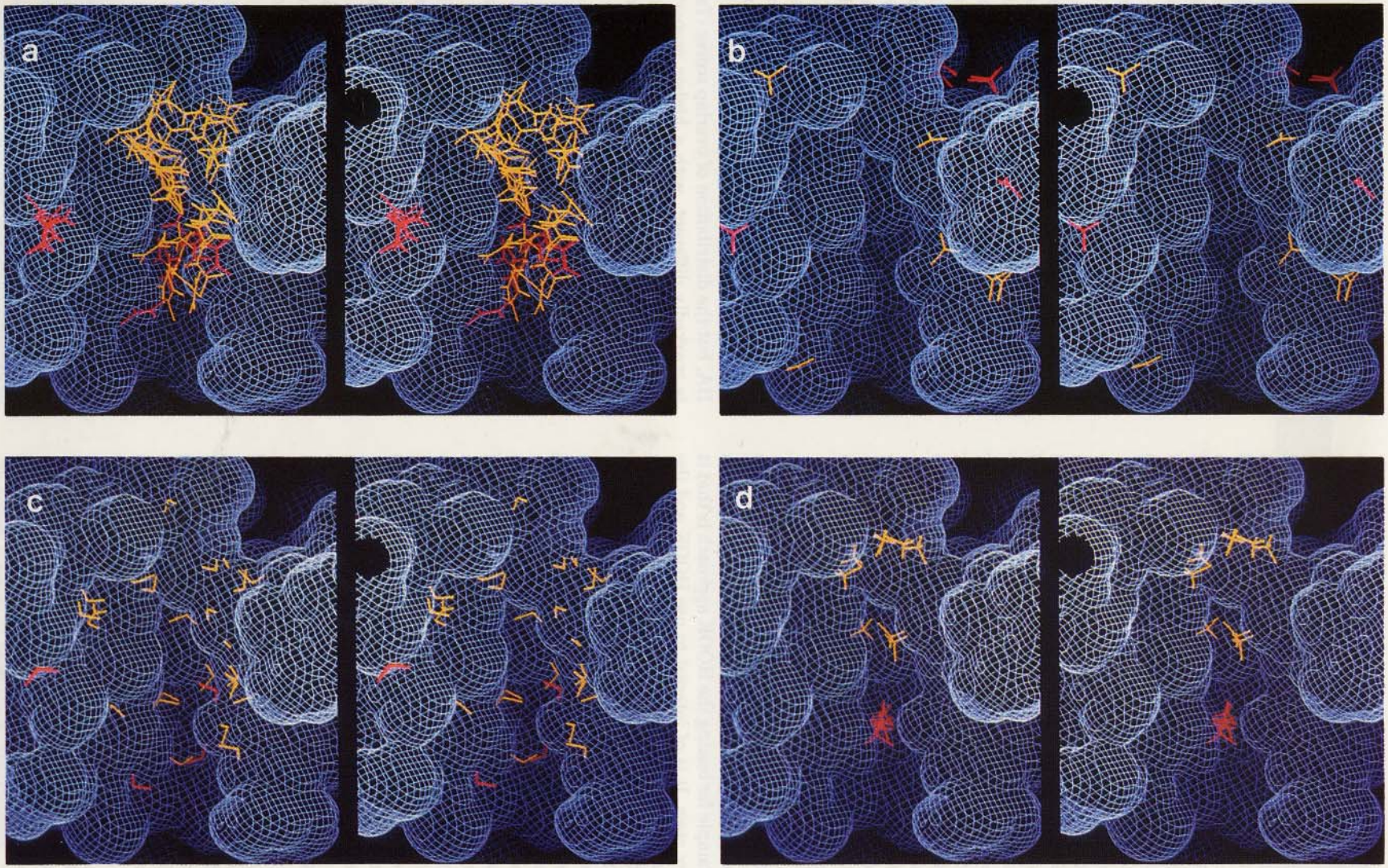


Fig. 7. The results of the MCSS calculations for the HA binding region. The molecular surface of HA computed on a 0.5-Å grid at a radius of 0.8 Å above the van der Waals surface of the molecule is shown in light blue, the primary functional group sites in red, and the secondary group site in yellow.

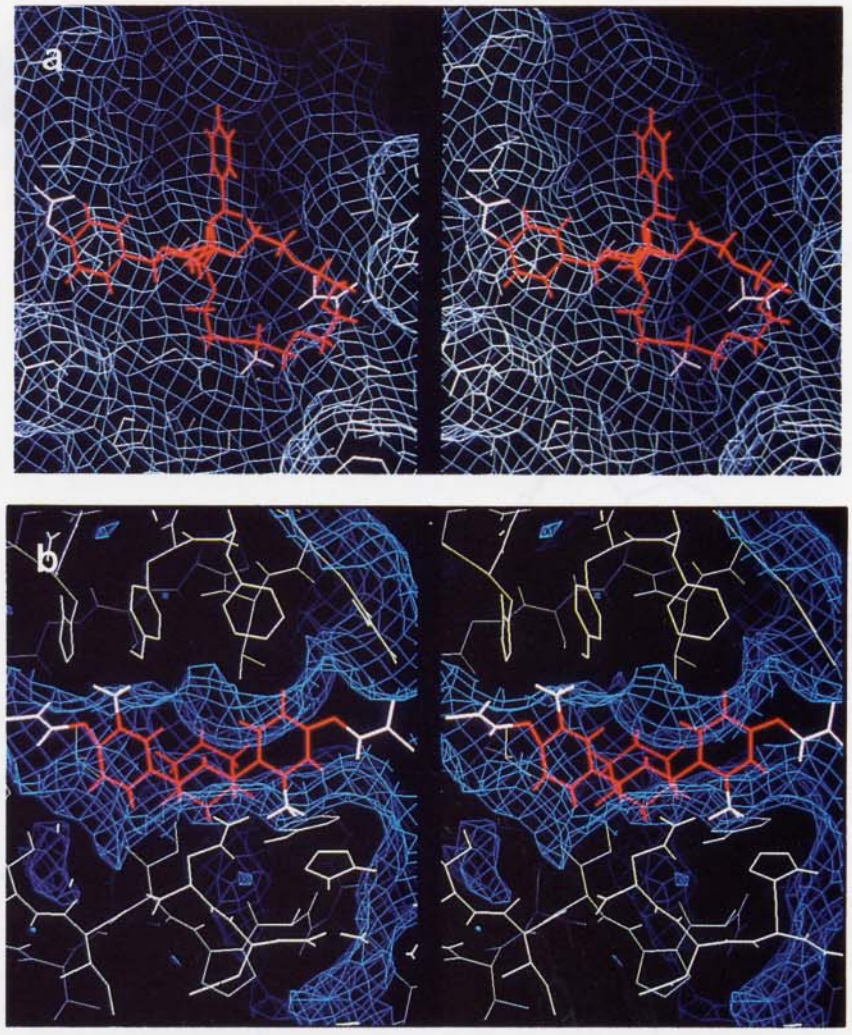


Fig. 8. (**a**) The binding region of HA with the molecular surface shown as a 1-Å grid computed as in Figure 6 (blue). The molecule in red is the hit with lowest MCSS interaction energy found by the HOOK program to link three functional group sites (shown in white). The molecule was constructed with a skeleton from database B. (**b**) The binding region of CAT with the molecule surface shown as a 1-Å grid computed as in Figure 6 (blue). The molecule in red is the hit with the lowest MCSS interaction energy found by the HOOK program to link four functional group sites (shown in white). The molecule was constructed with a skeleton from database B.

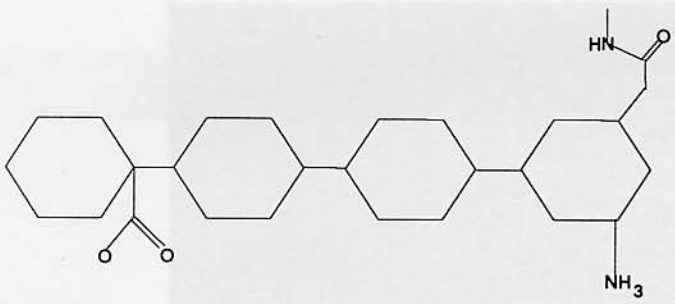
Important Elements of the HOOK Program

The HOOK program needs to cope with the combinatorial problems that come from screening large skeleton databases against multiple functional group sites. Examining all potential matches of pairs of functional group sites from a collection of 100 functional group sites with pairs of hooks from a skeleton database of 10,000 skeletons, each with 10 hooks, would require that 2.5×10^{12} such matches be considered. To make such searches computationally reasonable, the number of possible matches between pairs of hooks and pairs of sites that are considered must be made as small as possible, and each calculation has to be as rapid as possible. There are a number of features of HOOK that allow these two goals to be achieved.

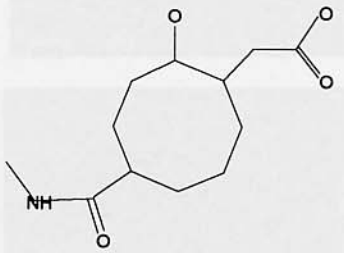
1. Before beginning the process of matching skeleton hooks and functional group sites, variables that will be used often (such as the distances, angles, and dihedral angles for the hook pair used in comparing internal geometries) are precomputed.

2. Since many of the functional group sites under consideration are overlapping and thus cannot be used together in a molecule, a functional group site pair exclusion array is precalculated to indicate which pairs of functional group sites should not be considered simultaneously.

3. Functional group sites are divided into primary sites and secondary sites on the basis of their interaction energy with the protein, and only pairs of primary sites are considered in the primary search. This restriction can eliminate many pairs of

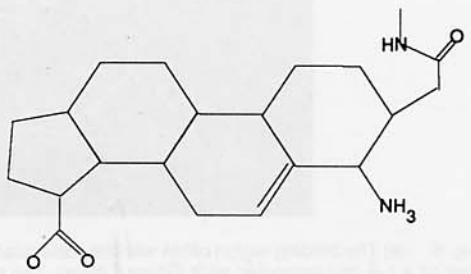


Molecule 12 Hit No: 3
Score: 129
MCSS: −178 kcal/mol

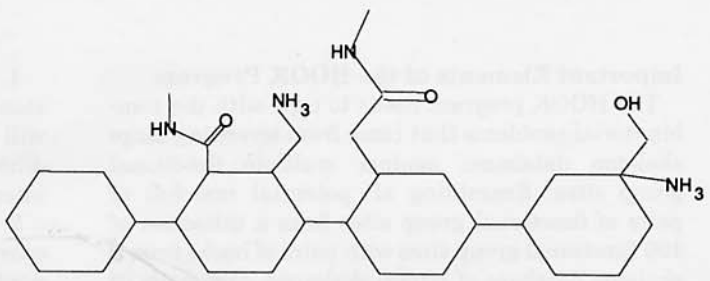


a

Molecule 13 Hit No: 8
Score: 62
MCSS: −148 kcal/mol



Molecule 14 Hit No: 3
Score: 91
MCSS: −159 kcal/mol



b

Molecule 15 Hit No: 4
Score: 133
MCSS: −157 kcal/mol

Fig. 9. Molecules constructed by HOOK from database B to satisfy the MCSS derived functional group sites within the (**a**) hemagglutinin and (**b**) CAT active site. The position in the hit list (as processed by the TABLE program and described in the text) and the overlap score, and the total MCSS interaction energy of the functional group sites with the binding region are given for each molecule.

sites from consideration in the primary search and guarantees that only molecules with at least two groups with strong interaction energy are produced as hits.

4. During each comparison of hook and site pairs, computationally expensive calculations (such as rms overlays and overlap score calculations) are performed only after the internal geometry screens. Since combinations that pass these screens are relatively rare, these time consuming calculations are performed relatively infrequently.

5. The number of combinations that do pass the internal geometry screens can be controlled by the tolerances on the distances, angles, and dihedral angles. Thus, the choice of these tolerances is important. Small tolerances restrict the number of combinations and improve the speed of the program, but they can limit the number and diversity of hits produced; large tolerances can lead to the opposite extreme. The values given in Figure 2 produced a reasonable number of hits for the HOOK runs presented in the present examples. Other choices may be appropriate for different systems.

6. HOOK can take advantage of multiprocessor computers by dividing a given HOOK job into separate jobs that each consider a subset of the skeleton database being used. The resulting output files can be merged.

The four HOOK calculations presented here (Table II) demonstrate how the computer time required varies with the nature of the binding region, the functional group sites and the skeleton databases. From a comparison of calculations (1) and (2), we see that increasing the number of primary functional group sites from 7 to 29 and the total number of sites from 7 to 106 increases the time for a HOOK calculation with the same skeleton database by just under a factor of two. This small increase is due to the fact that in calculation (1), which uses functional group sites derived from a known ligand, many of the pairs of functional group sites have the same relative position and orientation as pairs of skeleton hooks in the database. Hence, there are significantly more combinations of functional group site and skeleton hooks that pass the initial geometric test, and are taken forward for the calculation of least squares overlaps between the hooks and site vectors, overlap score and searches for additional groups. This is reflected by the large number of hits found in calculation (1). Comparing calculation (2) with calculation (3) it can be seen that even though database B contains more molecules than database A and there are many more skeleton hooks to be considered, the time for the calculation decreases. This is a result of the fact that molecules of database B have a much more varied architecture and thus there are substantially fewer combinations of pairs of functional group sites and pairs of skeleton hooks that satisfy the initial geometric criteria. This is also reflected in the number of hits.

Calculation (4) took a significantly longer time than the others because the overlap score was calculated considering all the atoms in the binding region (see Methods section). Making a subset, as was done for HA, would greatly reduce the computer time and still result in a satisfactory selection procedure.

We are currently constructing much larger databases of skeletons for use with HOOK from existing model and experimental structural databases. Examples are the complete Cambridge Crystallographic Databank and three-dimensional structures generated for the molecules in the Fine Chemicals Database (obtained from Molecular Design Limited). These databases will provide between 50,000 and 100,000 skeletons. If, on the basis of the examples above, we take the average time per skeleton to be 1 sec, a 100,000 skeleton database would take approximately 27 hr against a typical MCSS functionality map containing about 100 functional group sites. With increased computer power becoming available (multiple processors and higher speed individual processors) the required times are expected to be reduced significantly in the near future. The HOOK searches can clearly profit from parallelization at various levels; e.g., on a cluster a very simple approach would be to distribute the skeletons among the processors.

Many of the parameters, tolerances, and limits described in Figure 2 have significant effects on the characteristics of the molecules produced by HOOK, as well as the number of hits produced. The tolerances used in the internal geometry calculations control the type of linkages that can be achieved, as well as the variation from ideal geometry found in the generated molecules. The values for most of these variables were chosen so that the resulting molecules were realistic and representative, as well as not too numerous. More generous tolerances result in molecules with greater deviations from ideal geometry. This would introduce the requirement for geometry optimization in the postprocessing of molecules.

A distinctive feature of HOOK is its ability to use skeletons rather than existing molecules as the basis of potential ligands. Since the goal of HOOK is to identify rigid atomic linkages that can join a collection of functional group sites in many different ways, it is important to construct skeleton databases that sample a wide range of molecular architectures. The diversity of molecular architectures in these databases increases the probability that a linkage can be found for any given pair of functional groups sites, and helps generate linkages with different topological features to ensure that an optimal fit to the protein binding region can be achieved. This effect is similar to that of more generous geometry toler-

ances, but as skeleton molecules all have reasonable geometry, increased architectural variability is achieved without larger errors in geometry. Thus, more diverse skeleton databases allow for tighter geometry criteria.

Skeleton databases can be constructed either by systematically generating all chemically and structurally reasonable rigid skeletons, or by examining a large database of real molecules and extracting all rigid substructures to be used as skeletons. Both methods were used to generate the sample databases used in the examples presented in this paper. Generated databases can provide a wider range of architectures, but the generating procedure requires a method for regularizing geometries (e.g., minimizing the structures to obtain correct bond length and angles) and for removing unreasonable structures. Also, the synthetic accessibility of these structures needs to be assessed by comparison with available data. Since extracted databases contain only substructures of known molecules, this may compensate for their reduced structural diversity. The types of atoms used in these databases also have an effect on architectural diversity. We have found that while databases containing only carbon and hydrogen are most appropriate for HOOK in its present form, oxygens and nitrogens in rings and in bridges between rings greatly increase the diversity of skeletons. One possible procedure would be to construct skeletons with the polar atoms replaced by carbons. Alternatively, it would be necessary to extend the overlap score to include a simplified representation of electrostatic interactions. In the present implementation of HOOK, postprocessing of the candidate molecules is required to determine whether the polar groups in the skeletons make additional satisfactory interactions with the site.

Databases consisting of complete molecules can be useful in identifying candidates which satisfy the defined conditions for binding to a particular protein and they are likely to be available commercially or synthetically. This availability is an attractive feature of programs such as CLIX[16] or DOCK[17] that use such molecules in their searches. HOOK, too, can be used to perform such a search. This is done by using the HOOK facility that restricts individual hooks to accepting only one type of functional group. A database of known molecules can be processed to produce a skeleton database in which the only hooks are the bonds that join these functional groups to the molecule, labeling each hook with the type of functional group to which it is connected. HOOK can then use this database to search the functionality map as usual, except that a hook can be matched only with its associated type of functional group. Thus, if a hit is found, the known molecule can be oriented to fit the topology of the protein binding region with at least two of its functional groups in favorable positions. The number of molecules produced by such a search will be much smaller than in searches with skeletons, and the orientations are not as likely to produce favorable overlap scores. However, since the molecules are known to exist, problems of synthesis and availability are greatly reduced. At best they can be purchased from a known source, as is the case for the Fine Chemicals Database.

Another feature of HOOK is that the selection for topological compatibility between potential ligands and the protein binding region uses an overlap score that corresponds to an approximate van der Waals interaction energy calculation. The overlap score analyzes the fit of the potential ligand to the protein by examining all pairwise interactions between an atom in the ligand and an atom in the protein and modifying the score of the potential ligand based on the separation between the two atoms. Atom pairs whose separation is close to the optimal distance of the sum of their van der Waals radii score positively, and atom pairs that are significantly closer than this distance are penalized as their overlap increases. To make the calculation efficient, the function is divided into four regions (see Fig. 3d), and within each region the value of the function is linear with atomic separation. This greatly reduces the computation time needed for each evaluation, and the overall behavior of the function is still similar to the inverse Lennard–Jones function it was designed to emulate. The three variables that determine the boundaries of the linear regions, O_{cut}, O_{high}, and O_{low} were refined to correspond to van der Waals radii, and to give good results for a number of known ligands (see, for example, Fig. 5 and Table III). The main variable that changes the behavior of the overlap score is O_{scale}, which determines the magnitude of the penalty assessed for close contacts between the potential ligand and the protein binding region. Large negative values of O_{scale} penalize bad contacts significantly, even if the contact is slight, while low values allow for a few very bad contacts or a larger number of slightly bad contacts without greatly reducing the overlap score. This flexibility was included to offer the alternative of either disqualifying molecules if they make any bad contacts with the protein, or of assuming that a few bad contacts might be accommodated by small changes in the protein binding region. The typical value of -10 used in the examples presented here was chosen so that the HOOK hits with the highest overlap scores appeared, by visual inspection in QUANTA, to optimally fit the protein binding region. This choice is confirmed by the analysis of how the overlap score for sialic acid varies both as sialic acid is moved from its position in the crystal structure (Fig. 5) and how the overlap score for sialic acid varies in relation to other molecules as O_{scale} value is altered (Table III).

Different ligand design situations will require different criteria, so HOOK has been designed to pro-

duce thousands of candidate ligand molecules that satisfy the criteria set in a single calculation. The TABLE facility was developed within the HOOK program for extensive analysis of a large number of molecules to select the relevant molecules according to certain criteria. TABLE has spreadsheet like functions that allow users of HOOK to select molecules from the collection of HOOK hits on the basis of criteria that can be chosen and optimized interactively. Some of the features of the TABLE program have been illustrated in the examples presented in this paper. For example, the molecules shown in Figure 8a and b were selected to be the ones incorporating functional group sites with the lowest interaction energy and the highest overlap score per skeleton atom. The output from the HOOK program contains the essential information on a hit, such as the individual functional group details, the total overlap score, and the number of skeleton atoms. The TABLE program was used to manipulate these data by summing together all the functional group interaction energies, and by dividing the overlap score by the number of skeleton atoms. TABLE then sorted the list of hits on the basis of these two criteria. Using the simple spreadsheet functions, it is possible to sort the lists on a large number of different criteria. For example, it is straightforward to combine all the various measures of interaction energy and overlap score through some weighting scheme to produce a single score which can be used as a basis for sorting the hits. As experience with the HOOK program increases, it is probable that other criteria will be developed for postprocessing the candidate ligands. The flexibility of the TABLE program will be important for assessing different scoring regimes.

Applications of the HOOK Program

While HOOK was designed to use the output of MCSS, there are a number of other methods which can provide the required positions and orientations of functional group sites. Where structures of protein–ligand complexes are available from X-ray crystallographic NMR analysis, functional group positions and orientations can be identified. HOOK can then find novel molecular structures that link these sites in a different way from the original ligands. As demonstrated in the HA study, this serves to suggest molecules that are related to the original ligand but with increased rigidity or improved steric fit with the protein binding region. In addition, HOOK can also suggest new architectures of molecules that fit into the binding region and still satisfy the functional group requirements identified in the original ligand. When multiple ligand structures exist for a particular protein binding region, HOOK can use the functional group sites derived from these structures to generate potential hybrid ligands that fuse functional groups from different molecules. Such hybrids could combine aspects of several known ligands to create new classes of molecules with improved binding. It is also possible to use HOOK with skeleton databases constructed from known molecules. In this case, the program can be instructed to restrict the functional groups that can be attached to a given hook so that the candidate ligands correspond to known molecules (e.g., those in the Fine Chemicals Database, obtained from Molecular Design Limited).

Where multiple ligands are known to bind to a protein whose structure is not available, it is possible to build up a model of the receptor site[18] that may include a description of the excluded volume of the binding region. HOOK could find new molecules that contain some of the functional groups from these ligands and also complement the shape of the binding region. HOOK could also be used to identify novel molecules that link together flexible portions of proteins identified from an NMR analysis. The different conformations that result from NMR structure determination could define a number of functional group sites which can be linked together. In general, HOOK can be adapted to any set of information, provided the protein–ligand interactions can be expressed as functional group vectors within an excluded volume representing the binding region.

While by no means a distinctive feature of HOOK, it is important to note that HOOK treats the proteins, skeletons, and functional groups as rigid entities. The algorithm does not include any molecular dynamics or geometry optimization, as such calculations are very time consuming. This, however, does not mean that HOOK is limited in its ability to consider the flexibility in the data which describes the binding region, functional groups, or molecular skeletons used in the calculations. Flexibility of functional groups and of the protein can be taken into account when characterizing functional group sites (MCSS does not require that the functional groups and protein be rigid). Flexibility of skeleton molecules can be represented by including a number of conformations of each molecule in the skeleton database.

Comparison With Other Programs

MCSS represents a significant improvement over previous computational techniques for characterizing the chemistry of a protein binding region. Currently, the most widely used technique is that of Goodford, as implemented in the program GRID.[5,6] GRID calculates the interaction energy of functional groups represented as point with defined electrostatic and van der Waals properties of the protein represented on a grid. Contouring of this interaction energy allows for the definition of regions where various functional groups make favorable interac-

tions with the protein. Because functional groups are reduced to a point, no orientations or ideal positions of groups are computed in GRID. MCSS, on the other hand, positions and orients intact functional groups so that favorable interactions with the protein are optimized.

The advantages of the functional group sites as generated by MCSS can be seen when comparing HOOK to programs that use different methods for positioning functional groups. The program CLIX[16] uses the output of GRID to search the Cambridge Crystallographic Databank for small molecules that can position functional groups in the regions of interaction defined by GRID while simultaneously achieving reasonable steric fit with the protein binding region. The use of GRID to guide the placement of functional groups reduces the accuracy of the functional group positioning, and increases the complexity of the CLIX calculations. In addition, the absence of explicit functional group positions and orientations forces CLIX to rely on real molecules for its search templates, as there is no information available to guide the de novo creation of ligands.

The program CAVEAT[19] uses functional group vectors in a manner that is, in principle, similar to what is done in HOOK. However, CAVEAT is optimized to search through very large databases for molecules that link together just specified functional group vectors, and the search ignores the fit between the binding region and the molecule. Attempts were made to use CAVEAT to generate molecules starting with the MCSS results. It became clear that it was essential to have a way of searching databases using many functional group vectors and of selecting molecules that fitted into the binding region. The program HOOK was created to do these tasks.

The program DOCK[17] pioneered using geometric criteria to select ligands which best complement the shape of the protein binding region. Early versions of DOCK, however, did not consider the functionality of ligands, and thus integration with MCSS was not possible. More recent versions of DOCK incorporate chemical sense into the docking algorithm, but, like CLIX, use preexisting molecules.

The program LUDI[20,21] is closest in spirit to HOOK. However, rather than using MCSS minima, the program places a small number of "molecular fragments" in the protein binding region and uses a series of rules which define the fragments in relation to chemical constituents of amino acids. This is done on the basis of a set of heuristics derived from studies of the interactions found in known small molecule crystal structures. This type of placement is clearly restricted by the level of sophistication of the rules employed, and will in general be less accurate and less general than the placement of groups based on MCSS. By introducing such restrictions, LUDI achieves the useful feature of being able to do the required searches in real time so they can be used while working interactively.

Another approach creates molecules by allowing a potential ligand to "grow" in the active site, either by systematic build-up of the ligand (SPROUT[22] and GROW[23]) or by evolution of a set of ligands using genetic algorithms.[24] In this approach the programs attempt to satisfy both the chemical and steric requirements of the binding site at the same time. It is difficult to control the process to produce ligands that can be synthesized, although the types of ligands constructed are often suggestive of molecules that could readily be made.

In the approach described here, the ligand design problem is treated a step at a time so that chemical complementarity (functional groups from MCSS), steric complementarity (HOOK and overlap score), and synthetic considerations (databases of known molecular skeletons) can be considered in turn. Molecules suggested by the HOOK program could be optimized by procedures that modify or add to the molecule, such as the dynamic ligand design method.[25]

ACKNOWLEDGMENTS

We are grateful to Andrew Miranker, Peter Grootenhuis, Vincent van Geerestein, and Collin Stultz for stimulating discussions. Vincent van Geerestein at Organon, Inc. kindly supplied us with a codified data base of skeletons selected from the Cambridge Crystallographic Database. This was used in preliminary searches, but the work reported here made use of a set of skeletons obtained directly from the Cambridge Crystallographic Database. The research at Harvard was supported by the National Institutes of Health under Grant GM 39589 (Structural Foundations of Anti-Viral Drug Design). The work at York is supported by the SERC and the Wellcome Trust. M.B.E. was supported by a predoctoral fellowship from the National Science Foundation, and by Cindy Dillman.

REFERENCES

1. Ripka, W.C., Blaney, J.M. Computer graphics and molecular modeling in the analysis of synthetic targets. Topics Stereochem. 20:1, 1991.
2. Jolles, G., Woolridge, K.R.H., ed. "Drug Design: Fact or Fantasy?" London: Academic Press.
3. Dean, P.M. "Molecular Foundations of Drug-Receptor Interactions." Cambridge: Cambridge University Press, 1987.
4. Miranker, A., Karplus, M. Functionality maps of binding sites: A multiple copy simultaneous search method. Proteins: Struct. Funct. Genet. 11:29, 1991.
5. Goodford, P.J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. J. Med. Chem 28:849, 1985.
6. Bobbyer, D.A., Goodford, P.J., McWhinnie, P.M., Wade, R.C. New hydrogen-bond potentials for use in determining energetically favorable binding sites on molecules of known structure. J. Med. Chem. 32:1083, 1989.
7. Brünger, A.T., Karplus, M. Polar hydrogen positions in proteins: Empirical energy placement and neutron diffraction comparison. Proteins 4:148, 1988.

8. Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J. Comp. Chem. 4:187–217, 1983.
9. Fitzpatrick, P.A., Steinmetz, A.C., Ringe, D., Klibanov, A.M. Enzyme structure in a neat organic solvent. Proc. Natl. Acad. Sci. U.S.A. 90(18):8653–8657, 1993.
10. Allen, F.H., Bellard, S., Brice, M.D., Cartwright, B.A., Doubleday, A., Higgs, H., Hummelink, T., Hummelink-Peters, B.G., Kennard, O., Motherwell, W.D.W., Rodgers, J.R., Watson, D.G. The Cambridge Crystallographic Data Centre: Computer-based search, retrieval, analysis and display of information. Acta Crystallogr. B35:2331–2339, 1979.
11. Rusinko III, A., Skell, J.M., Baldwin, R., Pearlman, R.S. CONCORD. University of Texas at Austin, distributed by Tripos Associates, Inc., St. Louis, Missouri.
12. Wilson, I.A., Skehel, J.J., Wiley, D.C. Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 A resolution. Nature (London) 289:366–373, 1981.
13. Weis, W., Brown, J., Cusack, S., Paulson, J.C., Skehel, J.J., Wiley, D.C. Structure of the influenza-virus haemagglutinin complexed with its receptor, sialic acid. Nature (London) 333:426–431, 1988.
14. Leslie, A.G.W. Refined crystal structure of type III chloroamphenicol acetyltransferase at 1.75 Angstroms resolution. J. Mol. Biol. 213:167, 1990.
15. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. The Protein Data Bank: A computer-based archival file for macromolecular structures. J Mol. Biol. 112:535–542 (1977); Weng, J., Protein Data Bank. In: "Crystallographic Databases—Information Content, Software Systems, Scientific Applications." Allen, F.H., Bergerhoff, G., Sievers, R. eds. Bonn: Data Commission of the International Union of Crystallography, 1987: 107–132.
16. Lawrence, M.C., Davis, P.C. CLIX: A search algorithm for finding novel ligands capable of binding proteins of known three dimensional structure. Proteins 12:31–41, 1992.
17. DesJarlais, R.L., Sheridan, R.P., Seibel, G.L., Dixon, J.S., Kuntz, I.D., Venkataraghavan, R. Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure. J. Med. Chem. 28:849–857, 1988.
18. Marshall, G.R., Motoc, I. Approaches to the conformation of the drug bound to the receptor. In: "Topics in Molecular Pharmacology," Vol. 3. Burgen, A.S.V., Roberts, G.C.K., Tute, M.S. eds. Amsterdam: Elsevier, 1986: 115.
19. Bartlett, P.A., Shea, G.T., Telfer, S.J., Waterman, S. CAVEAT: A program to facilitate the structure-derived design of biologically active molecules. In: "Molecular Recognition: Chemical and Biological Problems." Vol. 78. Roberts, S.M., ed. Cambridge: Royal Society of Chemistry, 1989: 182.
20. Bohm, H.-J. The computer program LUDI: A new method for the de novo design of enzyme inhibitors. J. Comp.-Aided Mol. Design. 6:61, 1992.
21. Bohm, H-J. LUDI: Rule-based automatic design of new subsituents for enzyme inhibitor leads. J Comp.-Aided Mol. Design 6:593–606, 1992.
22. Moon, J.J., Howe, W.J. Computer design of bioactive molecules: A method for receptor based *de novo* ligand design. Proteins. Struc. Func. Genet. 11:314–328, 1991.
23. Gillet, V., Johnson, A.P., Mata, P., Sike, S., Williams, P. SPROUT: A program for structure generation. J. Comp.-Aided Mol. Design 7:127–153, 1993.
24. Payne, A.W.R., Glenn, R.C. Molecular recognition using a binary genetic search algorithm. J. Mol. Graphics 11:74–91, 1993.
25. Miranker, A., Karplus, M. Dynamic ligand design. In preparation.